Algorithmic statistics: normal objects and universal models

Alexey Milovanov
Moscow State University
almas239@gmail.com

December 7, 2015

Abstract

Kolmogorov suggested to measure quality of a statistical hypothesis (a model) P for a data x by two parameters: Kolmogorov complexity $C(P)$ of the hypothesis and the probability $P(x)$ of x with respect to P. The first parameter measures how simple is the hypothesis P and the second one how it fits. The paper [2] discovered a small class of models that are universal in the following sense. Each hypothesis S_{ij} from that class is identified by two integer parameters i, j and for every data x and for each complexity level α there is a hypothesis S_{ij} with $j \leqslant i \leqslant l(x)$ of complexity at most α that has almost the best fit among all hypotheses of complexity at most α. The hypothesis S_{ij} is identified by i and the leading $i - j$ bits of the binary representation of the number of strings of complexity at most i. On the other hand, the initial data x might be completely irrelevant to the the number of strings of complexity at most i. Thus S_{ij} seems to have some information irrelevant to the data, which undermines Kolmogorov's approach: the best hypotheses should not have irrelevant information.

To restrict the class of hypoteses for a data x to those that have only relevant information the paper [11] defined a notion of a strong model for x as a model P such that the total conditional complexity of P given x is negligible. An object x is called normal if for each complexity level α at least one its best fitting model of that complexity is strong. It was known that there are normal objects and there are non-normal objects. However, it was unknown how frequent are normal objects. In this paper we show that for every structure curve (representing the trade off between complexity of model and its fit) there is a normal string x that has that structure curve with $O(\sqrt{l(x)})$ accuracy. Our second result states that there is a normal object such that all its best fitting universal models are not strong for x. Our last result states that every best fit strong model for a normal object is again a normal object.

Keywords: algorithmic statistics, minimum description length, stochastic strings, total conditional complexity, sufficient statistic, denoising

1 Introduction

Let us recall the basic notion of algorithmic information theory and algorithmic statistics (see [8, 5, 10] for more details). As objects, we consider strings over the binary alphabet $\{0,1\}$. The set of all strings is denoted by $\{0,1\}^*$ and the length of a string x is denoted by $l(x)$. The empty string is denoted by Λ.

1.1 Algorithmic information theory

Let D be a partial computable function mapping pairs of strings to strings. *Conditional Kolmogorov complexity* with respect to D is defined as

$$C_D(x|y) = \min\{l(p) \mid D(p,y) = x\}.$$

In this context the function D is called a *description mode* or a *decompressor*. If $D(p,y) = x$ then p is called a *description of x conditional to y* or a *program mapping y to x*.

A decompressor D is called *universal* if for every other decompressor D' there is a string c such that $D'(p,y) = D(cp,y)$ for all p, y. By Solomonoff—Kolmogorov theorem universal decompressors exist. We pick arbitrary universal decompressor D and call $C_D(x|y)$ *the Kolmogorov complexity* of x conditional to y, and denote it by $C(x|y)$. Then we define the unconditional Kolmogorov complexity $C(x)$ of x as $C(x|\Lambda)$.

Kolmogorov complexity can be naturally extended to other finite objects (tuples of strings, finite sets of strings, etc.). We fix some computable bijection ("encoding") between these objects are binary strings and define the complexity of an object as the complexity of the corresponding binary string. It is easy to see that this definition is invariant (change of the encoding changes the complexity only by $O(1)$ additive term).

In particular, we fix some computable bijection between strings and finite subsets of $\{0,1\}^*$; the string that corresponds to a finite $A \subset \{0,1\}^*$ is denoted by $[A]$. Then we understand $C(A)$ as $C([A])$. Similarly, $C(x|A)$ and $C(A|x)$ are understood as $C(x|[A])$ and $C([A]|x)$, etc. By $\log n$ we denote binary logarithm.

Symmetry of information: $C(x) + C(y|x) \approx C(y) + C(x|y) \approx C(x,y)$. This equality holds with accuracy $O(\log(C(x)+C(y))$ and is due to Kolmogorov and Levin.

1.2 Algorithmic statistics: basic notions

Algorithmic statistics studies explanations of observed data that are suitable in the algorithmic sense: an explanation should be simple and capture all the algorithmically discoverable regularities in the data. The data is encoded, say, by a binary string x. In this paper we consider explanations (statistical hypotheses) of the form "x was drawn at random from a finite set A with uniform distribution". (As argued in [12], the class of general probability distributions reduces to the class of

Kolmogorov suggested in 1974 [4] to measure the quality of an explanation $A \ni x$ by two parameters: complexity $C(A)$ of A and the log-cardinality $\log|A|$ of A. The smaller is $C(A)$ the simpler is the explanation. The log-cardinality measures the *fit* of A—the lower is $|A|$ the more A fits as an explanation for any of its elements. For each complexity level m any model A for x with smallest $\log|A|$ among models of complexity at most m for x is called a *best fit hypothesis for* x. The trade off between $C(A)$ and $\log|A|$ is represented by the *profile* of x.

Definition 1.1. The *profile* of a string x is the set P_x consisting of all pairs (m, l) of natural numbers such that there exists a finite set $A \ni x$ with $C(A) \leqslant m$ and $\log_2|A| \leqslant l$.

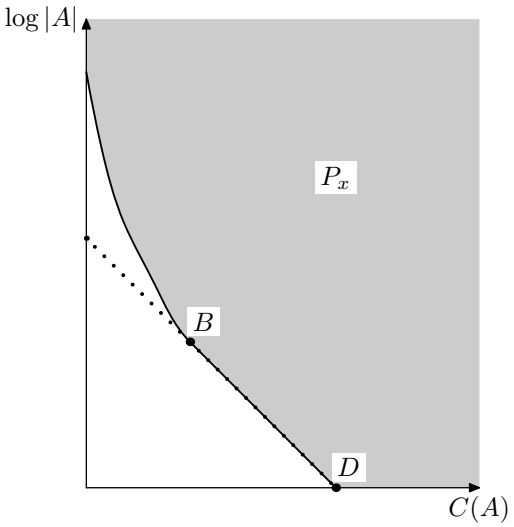

Figure 1: The profile P_x of a string x. {f1}

Both parameters $C(A)$ and $\log|A|$ cannot be very small simultaneously unless the string x has very small Kolmogorov complexity. Indeed, $C(A) + \log_2|A| \gtrsim C(x)$, since x can be specified by A and its index in A. A model $A \ni x$ is called *sufficient or optimal* if $C(A) + \log|A| \approx C(x)$. The value

$$\delta(x|A) = C(A) + \log|A| - C(x)$$

is called the *optimality deficiency* of A as a model for x. On Fig. 1 parameters of sufficient statistics lie on the segment BD. A sufficient statistic that has the minimal complexity is called *minimal* (MSS), its parameters are represented by the point B on Fig. 1.

Example 1.2. Consider a string $x \in \{0,1\}^{2n}$ such that leading n bits of x are equal zero, and the remaining bits are random, i. e. $C(x) \approx n$. Consider the model A for x that consists of strings $\{0,1\}^{2n}$ that n leading zeros. Then

$C(A) + \log |A| = \log n + O(1) + n \approx C(x)$, hence A is a sufficient statistic for x. As the complexity of A is negligible, A is a minimal sufficient statistic for x.

The string from this example has a sufficient statistic of negligible complexity. Such strings are called *stochastic*. Are there strings that have no sufficient statistics of negligible complexity? The positive to this question was obtained in [9]. Such strings are called *non-stochastic*. Moreover, under some natural constraints for every set P there is a string whose profile is close to P. The constraints are listed in the following theorem:

{th1}

Theorem 1.3. *Let x be a string of length n and complexity k. Then P_x has the following properties:*

1) $(k + O(\log n), 0) \in P_x$.

2) $(O(\log n), n) \in P_x$.

3) if $(a, b + c) \in P_x$ then $(a + b + O(\log n), c) \in P_x$.

4) if $(a, b) \in P_x$ then $a + b > k - O(\log n)$.

In other words, with logarithmic accuracy, the boundary of P_x contains a point $(0, a)$ with $a \leqslant l(x)$, contains the point $(C(x), 0)$, decreases with the slope at least -1 and lies above the line $C(A) + \log |A| = C(x)$. Coversely, given a curve with these property that has low complexity one can find a string x of length n and complexity about k such that the boundary of P_x is close to that curve:

{th2}

Theorem 1.4 ([12])**.** *Assume that we are given k, n and an upward closed set P of pairs of natural numbers such that $(0, n), (k, 0) \in P$, $(a, b + c) \in P \Rightarrow (a + c, b) \in P$ and $(a, b) \in P \Rightarrow a + b \geqslant k$. Then there is a string x of length n and complexity $k + O(\log n)$ whose profile is $C(P) + O(\log n)$-close to P. (We call subsets of $\mathbb{N}^2$ ε-*close *if each of them is in the ε-neighborhood of the other.) By $C(P)$ we denote the Kolmogorov complexity of the boundary of P, which is a finite object.*

1.3 Models of restricted type

{restrictedmodels}

It turns out that Theorems 1.3 and 1.4 remain valid (with smaller accuracy) even if we restrict the class of models under consideration to models from a class $\mathcal{A}$ provided the class $\mathcal{A}$ has the following properties.

(1) The family $\mathcal{A}$ is enumerable. This means that there exists an algorithm that prints elements of $\mathcal{A}$ as lists of strings, with some separators (saying where one element of $\mathcal{A}$ ends and another one begins).

(2) For every n the class $\mathcal{A}$ contains the set $\{0, 1\}^n$.

(3) The exists some polynomial p with the following property: for every $A \in \mathcal{A}$, for every natural n and for every natural $c < |A|$ the set of all n-bit strings in A can be covered by at most $p(n) \cdot |A|/c$ sets of cardinality at most c from $\mathcal{A}$.

Any family of finite sets sets of strings that satisfies these three conditions is called *acceptable*.

Let

$$P_x^{\mathcal{A}} = \{(a,b) \mid \exists A \ni x : A \in \mathcal{A}, C(A) \leqslant a, \log|A| \leqslant b\}.$$

Obviously $P_x^{\mathcal{A}} \subseteq P_x$. Let us fix any acceptable class $\mathcal{A}$ of models.

{th1a}

Theorem 1.5 ([13]). *Let x be a string of length n and complexity k. Then $P_x^{\mathcal{A}}$ has the following properties:*

1) $(k + O(\log n), 0) \in P_x^{\mathcal{A}}$.

2) $(O(\log n), n) \in P_x^{\mathcal{A}}$.

3) if $(a, b + c) \in P_x$ then $(a + b + O(\log n), c) \in P_x^{\mathcal{A}}$.

4) if $(a, b) \in P_x^{\mathcal{A}}$ then $a + b > k - O(\log n)$.

{th2a}

Theorem 1.6 ([13]). *Assume that we are given k, n and an upward closed set P of pairs of natural numbers such that $(0, n), (k, 0) \in P$, $(a, b + c) \in P \Rightarrow (a+c, b) \in P$ and $(a, b) \in P \Rightarrow a+b \geqslant k$. Then there is a string x of length n and complexity $k+O(\log n)$ such that* both sets *$P_x^{\mathcal{A}}$ and P_x are $C(P)+O(\sqrt{n \log n})$-close to P.*

Remark 1.7. Originally, the conclusion of Theorem 1.6 stated only that the set $P_x^{\mathcal{A}}$ is close to the given set P. However, as observed in [10], the proof from [13] shows also that P_x is close to P.

Notice that Theorems 1.5 and 1.6 are similar to Theorems 1.3 and 1.4. However the accuracy in Theorem 1.6 is worse than that in Theorem 1.4: $O(\sqrt{n \log n})$ in place of $O(\log n)$.

1.4 Sufficient statistics, denoising and useful information

Each sufficient statistic A provides a two-part code $y =$ (the shortest description of A, the index of x in A) for x whose total length is close to the complexity of x. The symmetry of information implies that $C(y|x) \approx C(y) + C(x|y) - C(x) \approx C(y) - C(x) \approx 0$. Thus x and y have almost the same information. That is, the two-part code y for x splits the information from x in two parts: the shortest description of A, the index of x in A. The second part of this two-part code is incompressible (random) conditional to the first part (as otherwise, the complexity of x would be smaller the the total length of y). Thus the second part of this two-part code can be considered as accidental information (noise) in the data x. In a sense every sufficient statistic A identifies about $C(x) - C(A)$ bits of accidental information in x. And thus any minimal sufficient statistic for x extracts almost all useful information from x.

1.5 Universal models

{um}

However, it turns out that this viewpoint is inconsistent with the existence of universal models, dicovered in [2]. Let L_m denote the list of strings of complexity at most m. Let p be an algorithm that, given m, enumerates all strings of L_m in some order. We assume that p is rather short , i. e. $l(p) = O(\log m)$. Denote by Ω_m the cardinality of L_m . Consider its binary representation, i. e., the sum:

$$\Omega_m = 2^{s_1} + 2^{s_2} + ... + 2^{s_t}, \text{ where } s_1 > s_2 > ... > s_t.$$

According to this decomposition and p, we split L_m into groups: first 2^{s_1} elements, next 2^{s_2} elements, etc. Let us denote by $S^p_{m,s}$ the group of size 2^s from the partition. Notice that $S^p_{m,s}$ is defined only for s that correspond to ones in the binary representation of Ω_m, so $m \geqslant s$.

If x is a string of complexity at most m, it belongs to some group $S^p_{m,s}$ and this group can be considered as a model for x. We may consider different values of m (starting from $C(x)$). In this way we get different models $S^p_{m,s}$ for the same x. The complexity of S^p_m is $m - s + O(\log m)$. Indeed, chop L_m into portions of size 2^s each, then $S^p_{m,s}$ is the last full portion and can be identified by m, s and the number of full portions, which is less than $\Omega_m/2^s < 2^{m-s+1}$. Thus if m is close to $C(x)$ and $x \in S^p_{m,s}$ then $S^p_{m,s}$ is a sufficient statistic for x. If p and q are two short algorithms enumerating all strings of complexity at most m then the families of sets $S^p_{m,s}$ and $S^q_{m,s}$ may be different. Nevertheless, there is a connect between them:

{difenum}

Theorem 1.8 ([10]). *Let p and q be two algorithms of size $O(\log m)$ that enumerate all strings of complexity at most m. Let x be a string that belongs to $S^p_{m,s}$. Then x belongs also to $S^q_{m,s+O(\log m)}$.*

This precision is suttisfied us, so we pick arbitary short enumerating algorithm p and denote $S^p_{m,s}$ by $S_{m,s}$.

The models $S_{m,s}$ are universal in the following sense:

Theorem 1.9 ([12]). [1] *Let A be any finite set of strings containing a string x of length n. Then there are $s \leqslant m \leqslant n + O(1)$ such that* {18}
1) $x \in S_{m,s}$,
2) $C(S_{m,s}|A) = O(\log n)$ (and hence $C(S_{m,s}) \leqslant C(A) + O(\log n)$),
3) $\delta(x|S_{m,s}) \leqslant \delta(x|A) + O(\log n)$.

In turns out that the model $S_{m,s}$ has the same information as the the number Ω_{m-s}:

{17}

Lemma 1.10 ([12]). *For every $a \leqslant b$ and for every $s \leqslant m$:*
1) $C(\Omega_a|\Omega_b) = O(\log b)$.
2) $C(\Omega_{m-s}|S_{m,s}) = O(\log m)$ and $C(S_{m,s}|\Omega_{m-s}) = O(\log m)$.
3) $C(\Omega_a) = a + O(\log a)$.

By Theorem 1.9 for every data x there is a minimal sufficient statistic for x of the form S_{ms}. Indeed, let A be any minimal sufficient statistic for x and let $S_{m,s}$ be any model for x that exists by Theorem 1.9 for this A. Then by item 3 the statistic $S_{m,s}$ is sufficient as well and by item 2 its complexity is also close to minimum. Moreover, since $C(S_{ms}|A)$ is negligible and $C(S_{m,s}) \approx C(A)$, by symmetry of information $C(A|S_{ms})$ is negligible as well. Thus A has the same information as S_{ms}), which has the same information as $C(\Omega_{m-s}|x)$ (Lemma 1.10(a)).

[1] This theorem was proved in [12, Theorem VIII.4] with accuracy $O(\max\{\log C(y) \mid y \in A\})$ instead of $O(\log n)$. Applying [12, Theorem VIII.4] to $A' = \{y \in A \mid l(y) = n\}$ we obtain the theorem in the present form

Thus if we agree that every minimal sufficient statistic extratcs all useful information from the data we must agree also that that information is the same as the information in the number of strings of compelxity at most i for some i.

1.6 Total conditinal complexity and strong models

The paper [11] suggests the following explanation to this paradox. Although conditional complexities $C(S_{m,s}|A)$ and $C(S_{m,s}|x)$ are small, the short programs that map A and x, respectively, to $S_{m,s}$ work in a huge time. A priori their work time is not bounded by any total computable function of their input. Thus it may happen that practicaly we are not able to find $S_{m,s}$ or Ω_{m-s} from a MSS A for x or from x itself.

Let us consider now programs whose work time is bounded by a total computable function for the input. We get the notion of *total conditional complexity* $CT(y|x)$, which is the length of the shortest total program that maps x to y. Total conditional complexity can be much greater than plain one, see for example [7]. Intuitively, good suffcient staistics A for x must have not only negligible conditinal complexity $C(A|x)$ (which follows from definition of a sufficient statistic) but also negligible *total* conditional complexity $CT(A|x)$. The paper [11] calls such models A *strong models for* x.

Is it true that for some x there is no MSS $S_{m,s}$ for x such that $CT(S_{m,s}|x)$ is negligible? The positive answer to this question was obtained in [11]: there are strings x whose all minimal sufficient statistics are not strong for x. Such strings are called *strange*. In particular, if $S_{m,s}$ is a MSS for strange string x then $CT(S_{m,s}|x)$ is large. However, a strange string has no strong MSS at all. An interesting question is whether there are strings x that do have strong MSS but have no strong MSS of the form $S_{m,s}$? This question was left open in [11]. In this paper we answer this question in positive. Moreover, we show that there is a "normal" string x that has no strong MSS of the form $S_{m,s}$ (Theorem 3.6). A string x is called *normal* if for every complexity level i there is a best fitting model A for x of complexity at most i (whose parameters thus lie on the border of the set P_x) that is strong. In particular, every normal string has a strong MSS.

Our second result answers yet another question asked in [11]. Assume that A is a strong MSS for a normal string x. Is it true that the code $[A]$ of A is a normal string itself? Our Theorem 4.5 states that this is indeed the case. Notice that by a result of [11] the profile $P_{[A]}$ of $[A]$ can be obtained from x by putting the origin in the point corresponding to parameters of A (see the point B on Fig. 1).

Then we show that there are normal strings with any given profile, under the same restrictions as in Theorem 1.3 (Section 2). And finally we provide tight bounds for number of strings with a given profile (Appendix) extending results of [14].

2 Normal strings with a given profile

{fixcurve}

In this section we prove an analogue of Theorem 1.4 for normal strings. We start with a rigorous definitions of strong models and normal strings.

A set $A \ni x$ is called *ε-strong statistic (model) for a string x* if $CT(A|x) < \varepsilon$. To represent the parameters of models for x consider the set P_x^ε representing the trade off between size and complexity of ε-strong models for x:

$$P_x^\varepsilon = \{(a,b) \mid \exists A \ni x : CT(A|x) \leqslant \varepsilon, C(A) \leqslant a, \log|A| \leqslant b\}.$$

It follows from the definition that $P_x^\varepsilon \subset P_x$ for all x, ε. Informally a string is called normal if for a negilgible ε we have $P_x \approx P_x^\varepsilon$. Formally, for integer parameters ε, δ we say that a string x is *ε, δ-normal* if $(a,b) \in P_x$ implies $(a+\delta, b+\delta) \in P_x^\varepsilon$ for all a, b. The smaller ε, δ are the stronger is the property of ε, δ-normality. The main result of this section shows that for some $\varepsilon, \delta = o(n)$ for every set P satisfying the assumptions of Theorem 1.3 there is an ε, δ-normal string of length n with $P_x \approx P$:

{VVfNS}

Theorem 2.1. *Assume that we are given an upward closed set P of pairs of natural numbers satisfying assumptions of Theorem 1.4. Then there is an $O(\log n), O(\sqrt{n \log n})$-normal string x of length n and complexity $k + O(\log n)$ whose profile P_x is $C(P) + O(\sqrt{n \log n})$-close to P.*

Proof. We will derive this theorem from Theorem 1.6. To this end consider the following family $\mathcal{B}$ of sets. A set B is in this family if it has the form

$$B = \{uv \mid v \in \{0,1\}^m\},$$

where u is an arbitrary binary string and m is an arbitrary natural number. Obviously, the family $\mathcal{B}$ is acceptable, that is, it satisfies the properties (1)–(3) from Section 1.3.

Note that for every x and for every $A \ni x$ from $\mathcal{B}$ the total complexity of A given x is $O(\log n)$. So $P_x^{\mathcal{B}} \subseteq P_x^{O(\log n)}$. By Theorem 1.6 there is a string x such that P_x and $P_x^{\mathcal{B}}$ are $C(P) + O(\sqrt{n \log n})$-close to P. Since $P_x^{\mathcal{B}} \subseteq P_x^{O(\log n)} \subseteq P_x$ we conclude that x is $O(\log n), O(\sqrt{n \log n})$-normal. □

The Theorem 4.7(2) provides an estimate of the number of normal strings with a given profile.

The proof of Theorem 2.1 is based on a technically difficult Theorem 1.6. However, for some sets P it can be shown direclty with a better accuracy of $O(\log n)$ in place of $O(\sqrt{n \log n})$). For instance, this happens for the smallest set P, satisfying the assumptions of Theorem 1.6, namely for the set

$$P = \{(m,l) \mid m \geqslant k, \text{ or } m + l \geqslant n\}.$$

Strings with such profile are called "antistochastic".

Definition 2.2. A string x of length n and complexity k is called *ε-antistochastic* if for all $(m,l) \in P_x$ either $m > k - \varepsilon$, or $m + l > n - \varepsilon$.

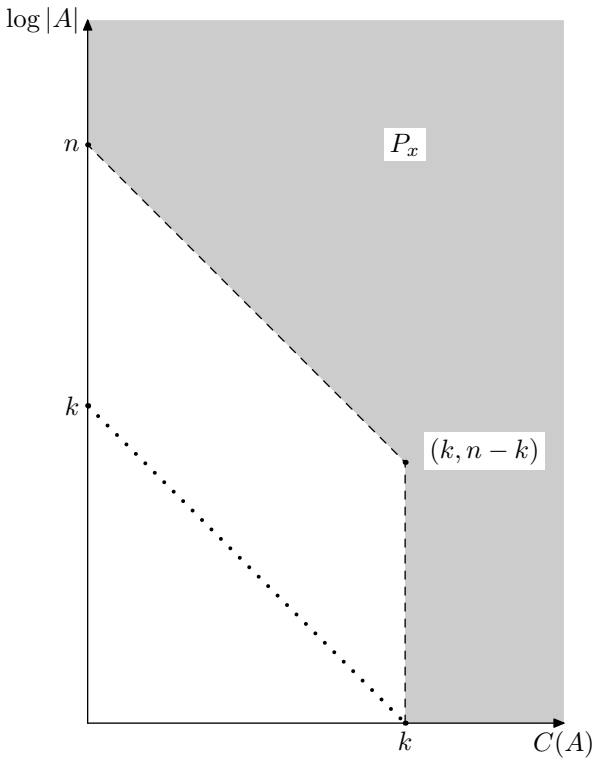

Figure 2: The profile of an ε-antistochastic string x for a small ε. {f2}

{c3}

Proposition 2.3. *For all n and all $k \leqslant n$ there is an $O(\log n)$-antistochastic string x of length n and complexity $k + O(\log n)$.*

This proposition can be derived from Theorem 1.4 or can be proved by a simple direct argument: it turns out that the lexicographically first string x of length n that is not covered by 2^{n-k}-element sets of complexity less than k satisfies Proposition 2.3.

Some interesting properties of antistochastic strings were established in [6]. It turns out that antistochastic strings are normal:

{pan}

Lemma 2.4. *Let x be an ε-anistochastic string of length n and complexity k. Then x is $O(\log n)$,$\varepsilon + O(\log n)$-normal.*

Proof. It suffices to show that for every $i \leqslant k$ there is a $O(\log n)$-strong statistics A_i for x with $C(A_i) \leqslant i + O(\log n)$ and $\log |A_i| = n - i$.

Let $A_k = \{x\}$ and for $i < k$ let A_i be the set of all strings of length n whose the first i bits are the same as those of x. By the construction $C(A_i) \leqslant i + O(\log n)$ and $\log |A_i| = n - i$. □

3 Normal strings without universal MSS

{without}

A model $A \ni x$ is called ε, δ-*good* for x if A is both ε-strong and δ-sufficient for x. A model A is called ε-*good* if A is ε, ε-good. Such models have the following remarkable property:

{thprg}

Theorem 3.1. *Let x be a string of length n and A be an ε-good statistic for x. Then for all $b \geqslant \log |A|$:*

$$(a, b) \in P_x \Leftrightarrow (a + O(\varepsilon + \log n), b - \log |A| + O(\varepsilon + \log n) \in P_{[A]}.$$

Proof. □

Lemma 3.3 states that if add to normal string some random bits then it remains normal. In the proof of this result we use the following theorem.

{mainpr}

Theorem 3.2 ([11])**.** *Let a, b and c be some integers and $x \in \{0,1\}^n$.*
Then $(a, b + c) \in P_x^{\varepsilon}$ implies $(a + b + O(\log n), c) \in P_x^{\varepsilon + O(\log n)}$.

(Thus, from every statistic we can constract another one, that is smaller but more complex.)

{stillnormal}

Lemma 3.3. *Let y be an ε-normal string of length n. Let z be a string of length m such that $C(z|x) \geqslant m - O(\log m)$. Then the string yz is $\varepsilon + \delta + O(\log(n+m))$-normal.*

Proof. Consider the set of all strings of length $n + m$ with the prefix y. Note that it is the good statistic for yz, so Theorem 3.1 shows relationships between sets P_y and P_{yz} (see Fig. 3). Since y is normal we need to prove the following two things: the first, if a point $(a, b) \in P_y^{\varepsilon}$ then a point $(a + m + O(\log(n + m)), b) \in P_{yz}^{\varepsilon + O(\log(n+m))}$, and the second one: for every integer $i \leqslant m$ the point $(C(y) + i + O(\log(n+m)), m - i) \in P_{yz}^{\varepsilon + O(\log(n+m))}$. Note that the second thing is the corolary from the first thing by Theorem 3.2.

To prove the first thing consider some ε-strong statistic A for y of complexity at most a and log-size at most b. Consider the set A_z that gets from A by the concatenation of all strings from A with z. It is the $\varepsilon + O(\log n + m)$-strong statistic for yz (because $CT(y|yz) = O(\log m)$) and its complexity is at most $a + m + O(\log(n + m))$ and log-size is at most b. □

Our main result of this section is Theorem 3.6 which states that there is a normal string x such that no set $S_{m,l}$ is not a strong MSS for x. In the proof we use the following rigorous definition of MSS from [11].

Definition 3.4. A set A is called a δ, ε, D-minimal sufficient statistic (MSS) for x if A is an ε-sufficient statisticc for x and there is no model B for x with $C(B) < C(A) - \delta$ and $C(B) + \log |B| - C(x) < \varepsilon + D \log C(x)$.

There is the following relataionships between MSS and sufficient statistics:

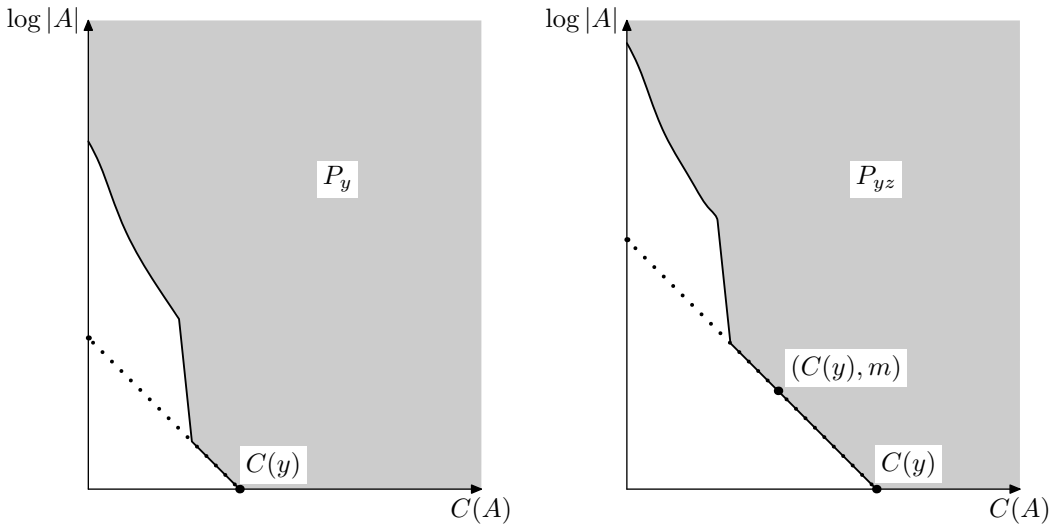

Figure 3: Profiles of strings y and yz from Lemma 3.3 {f3}

{MSSandSS}

Theorem 3.5 ([11], Theorem 13)**.** *For some constant D if B is ε-strong δ, ε, D-minimal sufficient statistic for x and A is an ε-sufficient statistic for x then $CT(B|A) = O(\varepsilon + \delta + \log C(x))$.*

To simplify notations in this section we call a model δ, ε-MSS if it δ, ε, D-MSS, where D is rather large paramater so that Theorem 3.5 holds. Now we are ready to formulate the main result of this section.

{nswnsmss}

Theorem 3.6. *For all large enough k there exist an $O(\log k)$-normal string x of complexity $3k + O(\log k)$ and length $4k$ such that:*

1) the profile P_x of x is $O(\log k)$-close to the gray set on Fig. 4.

2) There exists an α-strong α, α-MSS A for x, where $\alpha = O(\log k)$.

3) If $S^q_{m,l}$ is ε-strong δ, ε-MSS for x then $O(\varepsilon + \delta) + l(q) \geqslant k - O(\log k)$

Proof. Define x as a concatenation of strings y and z, where y is an antistochastic string of complexity k and length $2k$ and z is a string of length $2k$ such that $C(z|y) \geqslant 2k - O(\log k)$ (and so $C(x) = 3k + O(\log k)$). By Lemma 2.4 and Lemma 3.3 the string x is $O(\log k)$-normal. By Lemma 3.1 the profile of x is $O(\log k)$-close to the grey set on Fig. 4.

Consider the following set $A = \{yz' \mid l(z') = 2k\}$. From the shape of P_x it is clear that A is an $O(\log k)$-sufficent minimal statistic for x. Also it is clear that A is the $O(\log k)$-strong.

Let $S^q_{m,l}$ be an ε-strong δ, ε-MSS for x. We need to show that all three values ε, δ and $l(q)$ can not be small. By Theorem 3.5 we get $CT(S^q_{m,l}|A) = O(\varepsilon + \delta + \log k)$ and so $CT(s_0|y) = O(\varepsilon + \delta + \log k)$, where s_0 is the lexicorgaphic

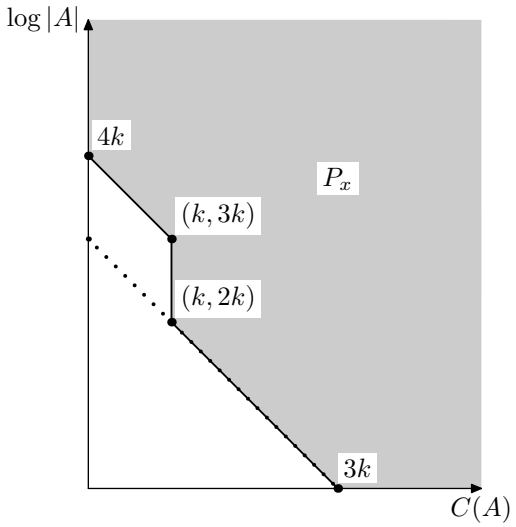

Figure 4: The profile P_x of a string x from Theorem 3.6. {f4}

last element in $S^q_{m,l}$. Denote by p a total program of length $O(\varepsilon+\delta+\log k)$ that produces s_0 on the input y. Consider the following set $B := \{p(y') \mid l(y') = 2k\}$. We claim that if ε and δ are not very big, then the complexity of any element from B is not greater than m. Indeed, if $\varepsilon + \delta \leqslant dk$ for enough small constant d, then $l(p) < k - O(\log k)$ and hence every element from B has complexity at most $C(B) + \log |B| + O(\log k) \leqslant 3k - O(\log k) \leqslant m$ (the last inequality holds because $x \in S^q_{m,l}$).

Let us enumerate all strings of complexity at most m by using the program q untill we meet all elements from B. Since $s_0 \in B$ there are at most 2^l elements of complexity m that we have not met. So, we can find the list of all strings of complexity at most m by using B, q and some l bits. Since this list has complexity at least $m - O(\log m)$ (because knowing this list and m we can compute a string of complexity more than m) we get: $C(B) + l(q) + l \geqslant m - O(\log m)$. Remember, that $m \geqslant 3k$, so: $O(\varepsilon+\delta+\log k)+l(q)+l \geqslant 3k-O(\log k)$. Since $S^q_{m,l}$ is the ε-sufficient statistic for x of log-size l, from the view of P_x it follows that $l \leqslant 2k + \varepsilon$ if $\varepsilon < k - O(\log k)$.

So we get: $O(\varepsilon + \delta) + l(q) \geqslant k - O(\log k)$ □

4 Hereditary of normality

{hereditary}

In this section we prove that every strong MSS for a normal string is normal too. We start with the following property of MSS.

{MSSOmega}

Lemma 4.1 ([11])**.** *For all large enough D the following holds:*
let A be a δ, ε, D-MSS for $x \in \{0,1\}^n$. Then $C(\Omega_{C(A)} | A) = O(\delta + \log n)$.

Proof. By Lemma 1.9 there is $S_{k,m} \ni x$ such that:

$$C(S_{k,m}|A) = O(\log n) \text{ and } \delta(x|S_{k,m}) \leqslant \delta(x|A) + O(\log n). \quad (1) \quad \{\texttt{Skm}\}$$

From $\delta(x|S_{k,m}) \leqslant \delta(x|A) + O(\log n)$ it follows that $S_{k,m}$ is an $\varepsilon + O(\log n)$-sufficient statistic for x. If D is rather large then $S_{k,m}$ is an $\varepsilon + D \cdot \log n$-sufficient statistic for x, hence, by definition of MSS:

$$C(S_{k,m}) > C(A) - \delta. \quad (2) \quad \{\texttt{estcs}\}$$

We can esstimate $C(\Omega_{C(A)}|A)$ as follows:

$$C(\Omega_{C(A)}|A) \leqslant C(\Omega_{C(A)}|\Omega_{C(S_{k,m})}) + C(\Omega_{C(S_{k,m})}|S_{k,m}) + C(S_{k,m}|A). \quad (3) \quad \{\texttt{en3terms}\}$$

To prove the statement of the lemma it remains to show that every term of the left part of the inequality above are equal to $O(\delta + \log n)$. For the third term it follows from (2).

To prove it for the first term note that $|C(A) - C(S_{k,m})| \leqslant \delta + O(\log n)$ by (2) and (3). Now the inequality $C(\Omega_{C(A)}|\Omega_{C(S_{k,m})}) \leqslant \delta + O(\log n)$ follows from the following simple lemma.

{ome}

Lemma 4.2. *Let a, b be some integers. Then*
$C(\Omega_a|\Omega_b) \leqslant |a - b| + O(\log(a + b))$.

Proof. Consider two cases. If $b \geqslant a$, then $C(\Omega_a|\Omega_b) = O(\log b)$ by the first statement of Lemma 1.10.

If $b < a$ we get $C(\Omega_b|\Omega_a) = O(\log a)$ by the same argument. By symmetry of information we get:

$$C(\Omega_a|\Omega_b) = C(\Omega_a) - C(\Omega_b) + O(\log(a + b)).$$

To conclude the required statement it remains to remember that $C(\Omega_a) = a + O(\log a)$ and $C(\Omega_b) = b + O(\log b)$ by Lemma 1.10. □

Now it remains to show that the second term of left part of the inequality (4) – $C(\Omega_{C(S_{k,m})}|S_{k,m})$ is equal to $O(\log n)$. But it is the easy corollary from the second and the third statements of Lemma 1.10. □

To simplify notation further we call a model δ, ε-MSS if it δ, ε, D-MSS, where D is a rather large parameter so that Lemma 4.1 holds.

Lemma 4.3 shows that every strong statistic A can be some modified to a strong statistic A' such that A' belongs to some simple partition.

A family of sets $\mathcal{A}$ is called *partition* if for every $A_1, A_2 \in \mathcal{A} : A_1 \cap A_2 \neq \varnothing \Rightarrow A_1 = A_2$. Note that for a finite partition we can define its complexity.

{part}

Lemma 4.3. *Let A be an ε-strong statistic for $x \in \{0,1\}^n$. Then there is a set A_1 and $\varepsilon + O(\log n)$-simple partition $\mathcal{A}$ such that:*
1) A' is $\varepsilon + O(\log n)$-strong statistic for x.
2) $CT(A|A') < \varepsilon + O(\log n)$ and $CT(A'|A) < \varepsilon + O(\log n)$.
3) $|A'| \leqslant |A|$.
4) $A' \in \mathcal{A}$.

Proof. A is an ε-strong statistic for $x \Rightarrow$ there is a total program p such that: $p(x) = A$ and $l(p) \leqslant \varepsilon$.

Now we will use the same construction as in Remark 1 in [11]. For every A denote by A' the following set: $\{x' \in A \mid p(x') = A, x' \in \{0,1\}^n\}$. Note that $CT(A'|A)$, $CT(A|A')$ and $CT(A'|x)$ are less than $l(p)+O(\log n) = \varepsilon + O(\log n)$ and $|A'| \leqslant |A|$.

If $x_1, x_2 \in \{0,1\}^n$ and $p(x_1) \neq p(x_2)$ then $p(x_1)' \cap p(x_2)' = \varnothing \Rightarrow \mathcal{A} := \{p(x)'|x \in \{0,1\}^n\}$ is a simple partition. □

By Lemma 1.9 for every $A \ni x$ there is a $B \ni x$, such that B is informational equalent $\Omega_{C(B)}$ and B is not worse model than A for x.

We state the same result for normal strings and for strong models.

{lch}

Lemma 4.4. *Let x be an ε, α-normal string with length n such that $\varepsilon \leqslant n$, $\alpha < \sqrt{n}/2$. Let A be an ε-strong statistic for x. Then there is a set H such that:*

1) H is an ε-strong statistic for x.

2) $\delta(x|H) \leqslant \delta(x|A) + O((\alpha + \log n) \cdot \sqrt{n})$ and $C(H) \leqslant C(A)$.

3) $C(H|\Omega_{C(H)}) = O(\sqrt{n})$.

Sketch of proof.

Consider the following sequence of statistics for x:

$$A_1 B_1 A_2 B_2 ...$$

Here $A_1 := A$, B_i is an improvement of A_i such that B_i is informational equalent to $\Omega_{C(B_i)}$. B_i exists by Lemma 1.9. A_{i+1} is a strong statistic for x that has a similar parameters as B_i (see Fig. 5). A_{i+1} exists because x is normal.

Denote by N the minimal integer such that $C(A_N) - C(B_N) \leqslant \sqrt{n}$. $C(A_{i+1}) \approx C(B_i)$ hence $N = O(\sqrt{n})$. Let $H := A_N$. By definition A_N (and H) is strong. From $N = O(\sqrt{n})$ it follows that the second condition is satisfied. From $C(A_N) - C(B_N) \leqslant \sqrt{n}$ and definition of B_N it is follows that the third condition is satisfied too (by using Lemma 4.2).

Proof. Let D be a statistic for x. Denote by $f(D)$ a statistic for x that is not worse than D and is equalent to Ω_t for some t, i. e. a statistic that exist for every D by Lemma 1.9: $C(f(D)|D) = O(\log n)$, $\delta(x|f(D)) \leqslant \delta(x|D) + O(\log n)$, $C(f(D)|\Omega_{C(f(D))}) = O(\log n)$ and $C(\Omega_{C(f(D))}|f(D)) = O(\log n)$.

Denote by $g(D)$ a statistic for x such that $C(g(D)) < C(D) + \alpha$, $\log |g(D)| < C(D) + \alpha$ and $g(D)$ is ε-strong. $g(D)$ exists for every D because x is ε, α-normal.

Consider the following sequence: $A_1 := A$, $B_1 = f(A_1)$, $A_2 = g(B_1)$, $B_2 = f(A_2)$ and so on.

Let us called a pair $A_i B_i$ a *big step* if $C(A_i) - C(B_i) > \sqrt{n}$. Denote by N the minimal integer such that $A_N B_N$ is not a big step. Let us show that $N = O(\sqrt{n})$. Indeed, $C(A_{i+1}) < C(B_i) + \alpha$, so: $C(A_{i+1}) - C(A_i) > \sqrt{n} - \alpha$ for every $i < N$. $C(A_1) \leqslant C(x) + CT(A_1|x) \leqslant n + \varepsilon$. So: $N \cdot (\sqrt{n} - \alpha) \leqslant n + \varepsilon$,

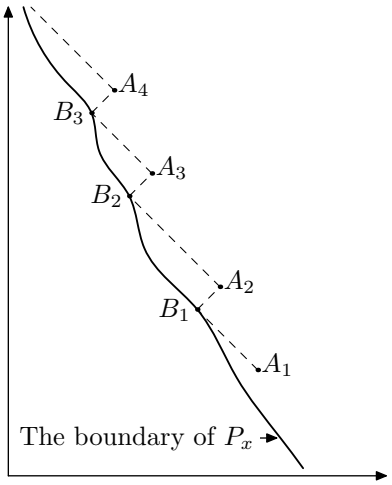

Figure 5: Parameters of statistics A_i and B_i {f5}

$\alpha < \sqrt{n}/2$, $\varepsilon \leqslant n$, hence $N = O(\sqrt{n})$. Define H as the follows: $H := A_N$. Let us show that A_N is satisfy all conditions:

1)A_N is an ε-strong by definition of g.

2) Let us make an estimate of $\delta(x|A_N)$. $\delta(x|A_{i+1}) \leqslant \delta(x|B_i) + 2 \cdot \alpha$, $\delta(x|B_i) \leqslant \delta(x|A_i) + O(\log n)$ for every i. So $\delta(x|A_N) \leqslant \delta(x|A_1) + N \cdot (2\alpha + O(\log n)) \leqslant \delta(x|A_1) + O(\alpha + \log n) \cdot \sqrt{n}$.

An estimate for complexity of A_N is simple too: $C(B_i) < C(A_i) - \sqrt{n}$ if $i < N$ and $C(A_{i+1}) < C(B_i) + \alpha$. $\alpha < \sqrt{n}/2 \Rightarrow C(A_{i+1}) < C(A_i)$ for $i < N$. Hence $C(A_N) \leqslant C(A_1)$.

3) To estimate $C(B_N|A_N)$ we use the following inequality:

$$C(A_N|\Omega_{C(A_N)}) \leqslant C(A_N|B_N) + C(B_N)|\Omega_{C(B_N)}) + C(\Omega_{C(B_N)}|\Omega_{C(A_N)}).$$

So, it remains to show that all terms in the right part are equal to $O(\sqrt{n})$.

It is true for the first term because $A_N B_N$ is not a big step. The second is term is equal to $O(\sqrt{n})$ by the definition of f.

An estimate for the third term gets from the inequality $|C(A_N) - C(B_N)| < \sqrt{n}$ (the pair $A_N B_N$ is not a big step and $C(B_N|A_N) = O(\log n)$) and Lemma 4.2. □

Remember, that a statistic is called ε-normal if its ε, ε-normal.

{hereditary_theorem}

Theorem 4.5. *Let A be both ε-strong and δ, ε-MGS for an ε-normal string x of length n. Assume that $\varepsilon \leqslant \sqrt{n}/2$. Then $[A]$ is $O((\varepsilon + \delta + \log n) \cdot \sqrt{n})$-normal.*

Sketch of proof:

We need to prove that $P_{[A]}$ is close to $P_{[A]}^{O((\varepsilon+\delta+\log n)\cdot\sqrt{n})}$. By Lemma 4.3 we can change A to A_1 that belongs to some simple partition $\mathcal{A}$.

Let $(a,b) \in P_{[A_1]}$. A_1 as A is a sufficient statistic for x hence by Theorem 3.1 $(a + \log|A_1|, b) \in P_x$ (see Fig. 6).

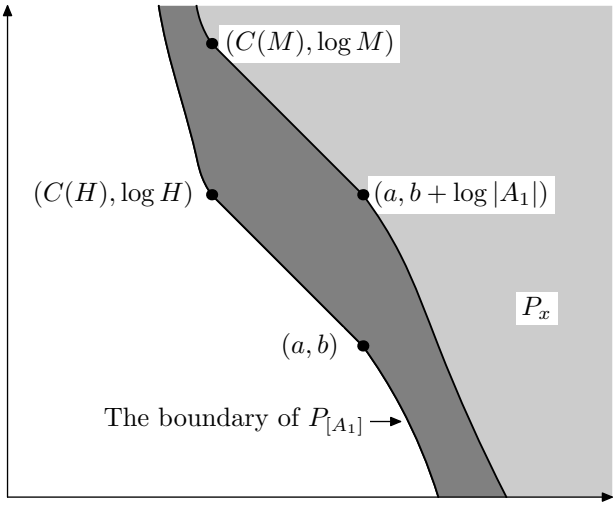

Figure 6: P_x is located $\log|A_1|$ higher than $P_{[A_1]}$ {f6}

x is normal, so a same point belongs to P_x^ε. By Lemma 4.4 we can improve this statistic to some statistic M such that M is informational equivalent to $\Omega_{C(M)}$. Again, we can change M to M_1 that belongs to some simple partition. Remember that by Lemma 4.1 A_1 is informational equivalent to $\Omega_{C(A_1)}$ too. So $C(M_1|A_1) \approx 0$ and from sufficiency of A_1 it follows that $\log|A_1| \approx \log|A_1 \cap M_1|$. M_1 belongs to simple partition ,

Consider the statistic H for A_1: all elements from $\mathcal{A}$ that have the same size of intersection with M_1 as A_1. H is a strong statistic for A_1 because $CT(M_1|A_1) \approx 0$. $C(H) \lesssim C(M_1)$ because $\mathcal{A}$ is simple, $\log|H| \leqslant \log|M_1| - \log|A_1|$ because $\log|A_1| \approx \log|A_1 \cap M_1|$.

Now note, that we can change H to a strong statistic with parameters closing to (a,b) by Theorem 3.2.

Proof. Now we give an accuracy proof of the theorem with respect of the notations in the sketch of proof.

Step 1: Change A to A_1.

A is ε-strong hence by Lemma 4.3 there is an $\varepsilon + O(\log n)$-strong statistic for x A_1 that belongs to $\varepsilon + O(\log n)$-simple partition $\mathcal{A}$ such that:

$$CT(A|A_1) < \varepsilon + O(\log n) \quad (4) \quad \{\texttt{total1}\}$$

$$CT(A_1|A) < \varepsilon + O(\log n) \quad (5) \quad \{\texttt{total2}\}$$

$$|A_1| \leqslant |A|. \quad (6) \quad \{\texttt{size}\}$$

We will show that $P_{[A_1]}$ is close to $P_{[A_1]}^{O((\varepsilon+\delta+\log n)\cdot\sqrt{n})}$ and after this we will prove the same result for A.

Let $(a,b) \in P_{[A]}$. We need to prove that there is a point that close to (a,b) that belongs to $P_{[A_1]}^{O((\varepsilon+\delta+\log n)\cdot\sqrt{n})}$. It is obviously if $a \geqslant C(A)$, so now and further we assume that $a < C(A)$. From (6) it is easy to see that $(a + O(\varepsilon + \log n), b + O(\varepsilon + \log n)) \in P_{[A_1]}$.

A is an ε-sufficient statistic for x. From this, (6) and (7) it follows that A_1 is an $O(\varepsilon + \log n)$-sufficient statistic for x.

Step 2: Construct M_1.

Now we will omit terms of kind $O((\varepsilon + \delta + \log n) \cdot \sqrt{n})$.

By Theorem 3.1 from sufficiency of A_1 it follows that $(a, b + \log |A_1|) \in P_x$. x is normal $\Rightarrow$ a point with similar parameters belongs to P_x^ε. A statistic with corresponding parameters can be improved by Lemma 4.4, i. e. there is an ε-strong statistic M for x such that:

$$C(M|\Omega_{C(M)}) = 0,\ C(M) \leqslant a, \tag{7}$$ {m111}

$$\text{and } \delta(x|M) \leqslant a + b + \log |A_1| - C(x). \tag{8}$$ {m113}

By Lemma 4.3 we can improve M to an $\varepsilon + O(\log n)$-strong statistic M_1 for x that belongs to $\varepsilon + O(\log n)$-simple partition $\mathcal{M}$:

$$CT(M|M_1) = 0,\ CT(M_1|M) = 0 \text{ and } |M_1| \leqslant |M|. \tag{9}$$ {M_1}

$$\text{From (8), (9) and (10) it follows that: } C(M_1|\Omega_{C(M_1)}) = 0, \tag{10}$$ {omega}

$$C(M_1) \leqslant a \tag{11}$$ {complexity}

$$\text{and } \delta(x|M_1) \leqslant a + b + \log |A_1| - C(x). \tag{12}$$ {optim}

Lemma 4.6. $\log |A_1 \cap M_1| = \log |A_1|$ *(up to $O((\varepsilon + \delta + \log n) \cdot \sqrt{n})$).*

Proof of Lemma. A is MSS, hence by Lemma 4.1 $C(\Omega_{C(A)}|A) = 0$.

$$C(A|A_1) = 0, \text{ so } C(\Omega_{C(A)}|A_1) = 0. \tag{13}$$ {omegaA}

Remember, that we assume $a < C(A)$. (12) states that $C(M_1) < a$, i. e. $C(M_1) < C(A)$. Hence, from Lemma 1.10 it follows that $C(\Omega_{C(M_1)}|\Omega_{C(A)}) = 0$. From this, (11) and (14) it follows that $C(M_1|A_1) = 0$. $C(M_1 \cap A_1) \leqslant C(A_1) + C(M_1|A_1)$, so:

$$C(M_1 \cap A_1) \leqslant C(A_1). \tag{14}$$ {cap}

A_1 is sufficient $\Rightarrow \log |A_1 \cap M_1| + C(M_1 \cap A_1) \geqslant C(A_1) + \log |A_1|$. From this and (15) it follows that $\log |A_1 \cap M_1| \geqslant \log |A_1|$. □

Step 3: Construct H.

Let us construct statistic for A. Denote by H all elements in $\mathcal{A}$ which has about the same size of intersection with M_1 as A_1, i. e.

$H = \{h \in \mathcal{A} | \lfloor \log h \cap M_1 \rfloor = \lfloor \log A_1 \cap M_1 \rfloor\}$.

$\mathcal{A}$ is partition $\Rightarrow |H| \leqslant |M_1|/(2 \cdot |A_1 \cap M_1|)$, so:

$$\log |H| \leqslant \log |M_1| - \log |A_1|.$$

We can get H by using M_1, $\mathcal{A}$ and $\lfloor \log A_1 \cap M_1 \rfloor$, so:

$$C(H) \leqslant C(M_1) + C(\mathcal{A}) = C(M_1).$$

By the same reason: $CT(H|A_1) \leqslant CT(M_1|A_1) + C(\mathcal{A}) + O(1) \leqslant CT(M_1|A_1)$.

To esstimate $CT(M_1|A_1)$ remember that $\mathcal{M}$ is partition, so there are not greater than $|A_1|/(2 \cdot |A_1 \cap M_1|)$ elements from $\mathcal{M}$ which log-size of intersection with A_1 is not less than $\lfloor \log M_1 \cap A_1 \rfloor$. So, we get:

$$CT(H|A_1) \leqslant CT(M_1|A_1) \leqslant \log |A_1| - \log |A_1 \cap M_1| + 1 + C(\mathcal{M}) = 0.$$

Thus H is strong statistic for A_1 of complexity atmost $C(M_1)$ and log-size atmost $\log |M_1| - \log |A_1|$:

$$(C(M_1), \log |M_1| - \log |A_1|) \in P^{O(\sqrt{n}+\varepsilon+\delta)}_{[A_1]}. \qquad (15) \quad \texttt{\{parH\}}$$

Step 4: Return to A.

(12) states that $a - C(M_1) \geqslant 0$. By Theorem 3.2 we can add this formula to the left part of (16) and substract it from the left part (i. e. make the statistic smaller but more complex):

$$(a, \log |M_1| - \log |A_1| - a + C(M_1)) \in P^{O(\sqrt{n}+\varepsilon+\delta)}_{[A_1]}.$$

By (13) the left part is not greater than b, i.e. $(a, b) \in P^{O(\sqrt{n}+\varepsilon+\delta)}_{[A_1]}$.

By the other words there is a set $B \ni [A_1]$ such that:

$C(B) = a$, $\log |B| = b$, $CT(B|[A_1]) = O(\sqrt{n} + \varepsilon + \delta)$. (5) states that there is a total programs p of length $\varepsilon + O(\log n)$ such that: $p([A_1]) = [A]$.

Consider the set $D := \{p(t) | t \in B\}$. $[A] \in D$, $\log |D| \leqslant \log |B|$, $C(D) \leqslant C(B) + l(p) + O(1) = a + O(\log n + \varepsilon)$, $CT(D|[A]) \leqslant CT(D|B) + CT(B|A_1) + CT(A_1|A) \leqslant O(\sqrt{n} + \varepsilon + \delta)$, hence:

$$(a + O(\log n + \varepsilon), b + O((\varepsilon + \delta + \log n) \cdot \sqrt{n})) \in P^{O(\sqrt{n}+\varepsilon+\delta)}_{[A]}.$$

□

Acknowledgments

The author is grateful to professor N. K. Vereshchagin for statements of questions, remarks and useful discussions.

References

[1] Bauwens, B., Makhlin, A., Vereshchagin, N., Zimand, M.: Short lists with short programs in short time. ECCC report TR13-007. http://eccc.hpi-web.de/report/ 2013/007/

[2] P. Gács, J. Tromp, P.M.B. Vitányi. Algorithmic statistics, *IEEE Trans. Inform. Th.*, 47:6 (2001), 2443–2463.

[3] Kolmogorov A. N.
"Three approaches to the quantitative definition of information". Problems Inform. Transmission, v. 1 (1965), no. 1, p. 1-7.

[4] A.N. Kolmogorov, Talk at the Information Theory Symposium in Tallinn, Estonia, 1974.

[5] Li M., Vitányi P., *An Introduction to Kolmogorov complexity and its applications*, 3rd ed., Springer, 2008 (1 ed., 1993; 2 ed., 1997), xxiii+790 pp. ISBN 978-0-387-49820-1.

[6] A. Milovanov *Some properties of antistochastic strings.* CSR 2015, LNCS 9139, pp. 1-11, 2015

[7] A. Shen, Game Arguments in Computability Theory and Algorithmic Information Theory. Proceedings of CiE 2012, 655–666.

[8] A. Shen, Around Kolmogorov complexity: basic notions and results. *Measures of Complexity. Festschrift for Alexey Chervonenkis.* Editors: V. Vovk, H. Papadoupoulos, A. Gammerman. Springer, 2015. ISBN: 978-3-319-21851-9

[9] A. Shen *The concept of (α, β)-stochasticity in the Kolmogorov sense, and its properties. Soviet Mathematics Doklady, 271(1):295–299, 1983*

[10] A. Shen, V. Uspensky, N. Vereshchagin *Kolmogorov complexity and algorithmic randomness.* MCCME, 2013 (Russian). English translation: http://www.lirmm.fr/~ashen/kolmbook-eng.pdf

[11] Nikolay Vereshchagin *"Algorithmic Minimal Sufficient Statistics: a New Approach". Theory of Computing Systems 56(2) 291-436 (2015)*

[12] N. Vereshchagin and P. Vitányi *"Kolmogorov's Structure Functions with an Application to the Foundations of Model Selection" . IEEE Transactions on Information Theory 50:12 (2004) 3265-3290. Preliminary version: Proc. 47th IEEE Symp. Found. Comput. Sci., 2002, 751–760.*

[13] Paul Vitányi, Nikolai Vereshchagin. *" On Algorithmic Rate-Distortion Function". Proc. of 2006 IEEE International Symposium on Information Theory Sunday, July 9 -Friday, July 14, 2006 Seattle, Washington.*

[14] V. V. V'yugin, On Nonstochastic Objects, *Probl. Peredachi Inf.*, 21:2 (1985), 3-9

Appendix

Here we deal with the following problem. Let P be some subset of $\mathbb{N}^2$. How many strings that has the profile that is close to P?

We assume that P suttisfy nessacary conditions of a profile, i. e. P is an upward close set such that $(a, b+c) \in P$ implies $(a+b, c) \in P$ for every integers a, b and c.

The number of strings which the profile is close to P is defined by the following parameters of P:

$$k_P = \min t | (t, 0) \in P;$$
$$m_P = \min t | (t, k_P - t) \in P;$$
$$n_P = \min t | (0, t) \in P.$$

So, if P coincide with the profile of some string x then k_P is equal to the complexity of x, n_P is equal to the length of x and k_P is equal to the complexity of a MSS of x.

{card}

Theorem 4.7. *1) There exist at least $2^{k_P - m_P - O(1)}$ strings which profile is $O(C(P) + \log n_P)$-close to P.*

2) There exist at least $2^{k_P - m_P - O(1)}$ strings which profile is $O(C(P) + \sqrt{n_P \log n_P})$-close to P that are $O(\log n_P)$, $O(\sqrt{n_P \log n_P})$-normal.

Proof. Consider a "bobtail" set $\tilde{P}$:

$$(a, b) \in \tilde{P} \Leftrightarrow (a + k_P - m_P, b) \in P.$$

Let us prove the first statement of the theorem. By Theorem 1.4 there is a string y of length $n_P + m_P - k_P + O(\log n_P)$ and complexity $m_P + O(\log n_P)$ which profile is $O(C(P) + \log n_P)$-close to $\tilde{P}$.

Let A be the set of all strings of length n_P with the prefix y. Denote by A' all strings in A such that has complexity $k_p - \log n_P$; the size of A' is at least $2^{k_P - m_P - O(1)}$. We claim that all strings in A' has the profile that close to P. Indeed, A is the good statistic for all elements in A', hence, as in Lemma 3.3 we get that the profile of all elements in A' is close to P.

The proof of the second statment is analogical. It is need to get a *normal* string y which profile is $O(C(P) + \sqrt{n_P \log n_P})$-close to $\tilde{P}$ – such string exists by Theorem 2.1. Now the proof is similar to the proof of the first statement. All strings from A' are normal by Lemma 3.3. □

Theorem 4.8 gets an upper estimate for the number of strings which profile is close to P. Denote by $L(P, \varepsilon)$ the set of all strings which profile is ε-close to P. We assume that ε is at most k_P.

To get an esstimate for $|L(P, \varepsilon)|$ we have to remember models $S_{k,l}$ from subsection 1.5. They have the following property:

$$C(S_{k,l}) + \log |S_{k,l}| \leqslant k + D \log k. \qquad (16) \quad \{\texttt{prSkl}\}$$

Here D is some constant. To formulate correct Theorem 4.8 we have to change the definition of m_P:

$$m_P(\varepsilon) := \min t | (t, k_P - t + \varepsilon + D \cdot (\log(k_P + \varepsilon))) \in P.$$

Here D is the constant from (17).

{uppest}

Theorem 4.8. $|L(P,\varepsilon)| \leqslant 2^{k_P - m_P(\varepsilon) + \varepsilon + O(\log k_P)}$.

To simplify the notation we take $m := m_P(\varepsilon)$. The theorem is follows from the following lemma.

{omp}

Lemma 4.9. *For every x from $L(P,\varepsilon)$:*
$C(\Omega_m | x) = O(\log n_P)$.

Proof of Theorem 4.8 from Lemma 4.9. Let x be some element from $L(P,\varepsilon)$. By the definitions of k_P and $L(P,\varepsilon)$ the point $(k_P + \varepsilon, 0)$ belongs to P_x, hence, $C(x) \leqslant k_P + \varepsilon$. By Lemma 1.10 $C(\Omega_m) = m + O(\log m)$. We claim that $C(x|\Omega_m)$ is not very large. Indeed, by symmetry of information $C(x|\Omega_m) = -C(\Omega_m) + C(x) + C(\Omega_m|x) + O(\log k_P)$. Using Lemma 4.9 we get $C(x|\Omega_m) \leqslant k_P - m + \varepsilon + O(\log k_P)$. But there are at most $2^{k_P - m + \varepsilon + O(\log k_P)}$ strings with such property. □

Proof of Lemma 4.9. Denote by k be the complexity of x. Let l be the ingteger such that x belongs to $S_{k,l}$. Since x belongs to $L(P,\varepsilon)$ we get $k \leqslant k_P + \varepsilon$. By (17) and the definition of m (i.e. of $m_P(\varepsilon)$) we get $C(S_{k,l}) \geqslant m$, hence, by Lemma1.10 we get $C(\Omega_m | S_{k,l}) = O(\log k_P)$. Also $C(S_{k,l}|x) = O(\log k_P)$ because $S_{k,l}$ is the sufficient statistic for x. So we get: $C(\Omega_m|x) \leqslant C(\Omega_m|S_{k,l}) + C(S_{k,l}|y) + O(\log n_P) = O(\log n_P)$. □